Problems and Possibilities of Cluster Analysis

With Application to Geochemical Data

Matthias Templ\textsuperscript{1}, Peter Filzmoser\textsuperscript{1} and Clemens Reimann\textsuperscript{2}

\textsuperscript{1}Vienna University of Technology

\textsuperscript{2}Norwegian Geological Survey

\textit{Bern}

10. September 2007
Stability of Cluster Analysis

For real data sets without obvious grouping structure the results of cluster analysis depends on:

1. Input data - the selection of variables

2. Preparation of the data

3. Distance measure used

4. Clustering method

5. Number of clusters

Changing one parameter may result in completely different cluster results.
1. Input Data - Variable Selection

```r
> library(mvoutlier)
> library(clustTool)
> data(humus)
> a <- agnes(t(prepare(humus[, -c(1:3)])))
> plot(a, which.plots = 2, col = c(4), col.main = 3, col.sub = 2)
```
1. Input Data - Variable Selection

```r
> library(mvoutlier)
> data(humus)
> a <- agnes(t(prepare(humus[, -c(1:3)])))
> plot(a, which.plots = 2, col = c(4), col.main = 3, col.sub = 2)
```

A chemical process can be seen in more detail in a map (later) by choosing similar variables.
A selection of variables may be useful when clustering high-dimensional data because . . .

- clustering with all variables may hide underlying processes
- we want to see some processes in more detail
- the inclusion of one irrelevant variable may hide the real clusters in the data

One easy way (amongst others) for variable selection can be done by graphical inspection of a dendrogram which results from hierarchical clustering of variables.
2. Data Preparation

Most of real data in practice can have some or all of these properties:

- neither normal nor log-normal, maybe compositional
- strongly skewed
- often multi-modal distributions
- weak clustering structures
- data include outliers
- variables show a striking difference in the amount of variability
2. Data Preparation

If a good clustering structure exists for a variable we expect a distribution with two or more modes. A transformation (e.g. with a box-cox transformation) will preserve the modes but remove large skewness.

If the data are of *compositional* nature, i.e. the observations sum up to a constant, we must *open* the data with a suitable *logratio* transformation.

- additive logratio transformation (alr)

  \[ x^{(j)} = (x_1^{(j)}, \ldots, x_{D-1}^{(j)})' = \left( \log \frac{x_1}{x_j}, \ldots, \log \frac{x_{j-1}}{x_j}, \log \frac{x_{j+1}}{x_j}, \ldots, \log \frac{x_D}{x_j} \right)' \]

- centered logratio transformation (clr)

  \[ y^{(j)} = (y_1^{(j)}, \ldots, y_{D-1}^{(j)})' = \left( \log \frac{x_1}{\sqrt{\prod_{i=1}^D x_i}}, \ldots, \log \frac{x_D}{\sqrt{\prod_{i=1}^D x_i}} \right)' \]

- isometric logratio transformation (ilr)

  \[ y^{(j)} = (y_1^{(j)}, \ldots, y_{D-1}^{(j)})', z_i = \sqrt{\frac{i}{i+1}} \log \sqrt{\frac{\prod_{j=1}^i x_j}{x_{i+1}}} \]

The results will change if the transformation will be changed.
2. Data Preparation

**Standardisation** of the variables is needed if the variables show a striking difference in the amount of variability.

**Outliers** can influence the clustering (depends on which clustering algorithms is chosen).

Removing outliers before clustering may be useful.

Finding outliers is not a trivial task, especially in high dimensions. (you can do this e.g. with Package *mvoutlier* from Filzmoser et al. (2005))
3. Distance Measure

- Almost all clustering methods are applied on a distance matrix, received from distances between observations using a certain distance measures like Euclidean distance, Manhattan distance or other.

- The results depend on the choice of the distance measure.
3. Distance Measure

Comparing clustered data (clustered with Average Linkage) and clustered bootstrap samples of the data with the Rand Index (with 6 clusters). The Rand Index accounts in a new clustering for observations which belong to the same cluster as in the original partition.
4. Clustering Method

E.g. in R there are a lot of cluster methods available, but only few are suited for complex data without clear clustering structure, depending on the distance measure and transformation used.

Using a geochemical data set with known clustering structure (OSLO data, 9 clusters) we get following amount of missclassifications (out of 350 observations):

<table>
<thead>
<tr>
<th></th>
<th>Ward</th>
<th>kmeans (best results after 10 starts)</th>
<th>Mclust</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Eucl</td>
<td>Manh</td>
<td>Gower</td>
</tr>
<tr>
<td>no</td>
<td>30</td>
<td>11</td>
<td>7</td>
</tr>
<tr>
<td>log</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>clr</td>
<td>23</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>alr</td>
<td>17</td>
<td>22</td>
<td>4</td>
</tr>
<tr>
<td>ilr</td>
<td>14</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
5. Number of Clusters
5. Number of Clusters
Global cluster validity measures which evaluates all clusters at once are, in general, not ideal suited to evaluate cluster results and to find a good set of variables, the optimal transformation/distance measure or the optimal number of clusters in complex data.

Local cluster validity measures which evaluates each cluster separately are useful.

A lot of parameters can influence the results.

We propose an explorative way of doing cluster analysis when analysing complex data.

The developed R-package *clustTool* does allow such explorative way and gives cluster analysis in the hand of the users.
• We extensively use the object-oriented class system of R.
• We developed a simple GUI using tcltk.
• Plots will be shown and results will be available in the R workspace.
Pollution

Highest pollution visualised by cluster 9
This can be seen in the graphic on the right
e.g. Co, Cu, Ni typical elements for reflecting pollution

Mclust on scaled and transformed humus data
Validity measure on each cluster
cluster size
Visualising all clusters each in an own map

Seaspray
Cluster 5
(Greyscale depends on validity measure in each cluster)
Fuzzy Clustering
Demo

...if there is time left.
Conclusions

- Applying cluster analysis on real data results in highly non-stable results for many reasons.

- The selection of variables, the selection of the transformation, the selection of the distance measure and the selection of the optimal number of clusters is a non-trivial task when using real data.

- Cluster analysis can be seen as explorative data analysis to get ideas about your data.

- Interactive tools are very helpful.